植物物种多样性语义知识抽取研究*

刘建华[1,2]　王　颖[1]　张智雄[1]　李传席[3]

[1](中国科学院文献情报中心　北京 100190)

[2](中国科学院大学　北京 100049)

[3](中国长城资产管理股份有限公司　北京 100045)

摘要: 【目的】拓展以物种为中心的植物物种多样性抽取框架, 探索实现语义知识抽取方法。【方法】结合当前生物多样性抽取的主流研究, 以物种为中心, 设计包含多种实体及实体间关系的知识抽取框架, 利用已有的众多专业数据库, 设计并实现相应的识别方法。【结果】设计以物种为核心的知识抽取框架, 探索实现多种实体及实体间关系的语义知识抽取方法, 拓展植物物种多样性领域抽取内容和思路。【局限】实体识别的完整性和准确性受底层知识库影响较大, 且实体间关系的类型局限于共现、上下位类、语法关系几类, 还需进一步研究。【结论】本研究拓展了植物物种多样性抽取内容和思路, 可有效支持语义检索、科学计算。

关键词: 植物物种多样性　植物物种　知识抽取　关系识别

分类号: G250

1　引　言

当前, 气候变化、自然灾害等原因导致物种灭绝速度越来越快, 针对生物多样性保护与持续利用的研究日益成为生物多样性研究的焦点, 植物物种作为生物多样性领域的重要内容之一, 针对其多样性的研究也吸引了众多科研人员。如何帮助科研人员从大量富含植物物种名称、基因、实验设备等实体的文档中快速发现所需信息, 是植物物种多样性信息学面临的重要问题之一。针对此, 越来越多的研究者正努力尝试利用现有众多的植物物种多样性专业数据库, 如物种名录、标本库、图片库、基因库等, 从植物物种多样性描述文本或文献中提取知识对象, 并借助语义内容标注技术实现自动深层标引, 实现数字资源之间的语义集成和关联, 从而为进一步的语义检索、数据挖掘、科学计算提供支撑。

本文在当前植物物种多样性信息抽取领域相关研究的基础上, 结合中国科学院文献情报中心"建设生物多样性领域本体构建与语义组织应用示范平台"的实际要求, 以植物物种多样性为目标领域, 设计了植物物种多样性语义知识组织框架, 探索实现了针对框架中定义的语义知识单元的抽取方法, 开发了相应的植物物种多样性示范平台。

2　相关研究概述

在众多研究者的努力下, 目前已经有不少针对生物多样性领域的信息抽取工具, 这些工具或者采用单一的自然语言处理、词典、机器学习、规则模板、浅度或深度句法解析等方法, 或者融合上述几种方法进行识别, 识别的内容多数集中于物种的各类名称(科学命名、别名、俗名、变种名等), 部分工具涉及对物种的性状的识别。Thessen 等[1]综述了当前在生物多样性领域使用自然语言处理和机器学习算法实现物种名称识别的相关研究; Naderi 等[2]介绍了 GATE 框架下提供

通讯作者: 刘建华, ORCID: 0000-0002-4003-8834, E-mail: liujh@mail.las.ac.cn。

*本文系国家"十二五"科技支撑计划项目"面向外文科技文献信息的知识组织体系建设与应用示范(STKOS)"的子课题"信息资源自动处理、智能检索与 STKOS 应用服务集成"(项目编号: 2011BAH10B05)和国家自然科学基金项目"基于语言网络的文本主题中心度计算"(项目编号: 61075047)的研究成果之一。

的生物医学领域的各种抽取工具。上述文献对常规的生物多样性信息抽取流程，主流的信息抽取方法进行了全面的评述，并对各个阶段的主要信息抽取工具进行了全面的综述。本文不再对上述内容进行重复梳理，而是结合当前一些重要的生物多样性信息抽取工具，重点对植物物种多样性领域抽取的内容进行探讨，希望在此基础上对笔者提出植物物种多样性知识抽取框架提供参考支撑。

目前植物物种多样性的抽取研究，主要内容可以归纳为以下几个方面:

2.1 物种名称识别及规范

由于语种、地方称谓等差异，科技文献中出现的同一个物种名称是多种多样的。有的是标准规范的双名制命名法(或三名制命名法)形成的拉丁文名，即属名加种名，且属名在前，种名在后，属名第一个字大写，种名小写，属种名称均为全称，后面通常还会跟随着物种命名人的姓氏[3]；有的取属名首字母、种名全称的缩写方式；有的会采用物种的俗名(可能是英文，也可能是其他语种，同一个物种在不同的国家或地区也可能会有不同的俗名)[4]。这些问题的存在大大增加了物种名称识别的难度。因此目前有不少研究者专门针对物种名称的识别、规范及组织进行研究，这也是当前植物物种多样性抽取相关研究的主流。这些研究成果中比较典型的包括可用作物种名称识别与规范词典的 NCBI Taxonomy[5]、BioNames[6](一个将动物名称与其来源描述、分类及进化树关联的在线数据库)、物种 2000 全球生物物种名录[7]，也包括各种比较成熟物种名称识别工具如 NetiNeti[8]、OrganismTagger[9]、Linnaeus[4]、TaxonGrab[10]等。

2.2 物种性状识别

对物种分类学研究人员而言，物种的各类性状描述信息，如根、茎、叶的颜色、长度等，是界定物种门类的重要参考信息。因此，有一部分生物信息学研究人员着力于探索物种各类性状的自动识别方法。Taylor[11]在分析文本语法特征的基础上，以人工方式建立规则和词典，实现了物种部位、特征及状态等描述信息的识别。Tang 等[12]在相关研究基础上，通过预定义模板的方式，有监督地学习生成相关的规则，实现了对物种叶子的形状、大小、颜色、排列及果实的形状特征的识别。CharaParser 采用启发式方法和句法特征生成规则，较好地实现了对物种多类性状的识别[13]。段宇锋等[14]持续探索着中文植物物种多样性描述文本中形态信息的抽取。

2.3 生物网络识别

各种生物实体(物种、分子、基因、蛋白等)之间存在多种关系，这些关系可以用网络图的方式表达出来，进而通过对图的分析实现对生物系统的分析[15-16]。蛋白质和基因是生物医药领域普遍关注的重点内容，关于这类知识的识别研究并不限在植物物种多样性领域开展。当前植物物种多样性相关的文献中，可通过对物种基因测序的方式来鉴定物种的亲缘性，也可通过采用蛋白质或基因技术影响或改变生物的内外环境或特征，从而研究相关问题。因此，对蛋白质和基因的识别更多地不仅仅是识别出蛋白质、基因等命名实体，而是识别出各类生物实体之间通过动词(或动词短语)、介词(或介词短语)、所有格等关联而成的生物网络关系。

综观目前生物多样性领域，尤其是植物物种多样性的信息抽取研究内容，多数是围绕某一类信息进行识别方法的探索，并以结构化描述植物物种的多样性特征或辅助判别物种为最终目标，鲜有面向科技文献内容的知识化组织和语义检索的系统研究与框架设计。本文基于当前研究成果，进一步系统化设计相应的语义组织知识框架，并从实际应用的角度探索相应知识单元的快速识别方法。

3 语义知识框架设计

要开展植物物种多样性的语义知识抽取工作，首先要明确需要从目标资源中抽取哪些内容，即要构建合理的语义知识描述框架，该框架是描述本领域需要抽取的语义知识单元及关联关系的重要依据，也是后续知识组织揭示的重要支撑。因此，在对现有相关研究分析的基础上，结合中国科学院文献情报中心“建设生物多样性领域本体构建与语义组织应用示范平台”的实际要求，设计了支撑该示范平台的语义知识框架。本节将详细介绍该框架的设计流程与框架内容。

3.1 语义知识框架的层级

在构建本框架的过程中，首先以“Oryza sativa(水稻物种)”为检索词，在 PubMed 数据库的 Plant Physiology、The Plant Cell 期刊上实施检索，并

从检索集合中随机选择 100 篇科技文献进行人工标引，再通过咨询中国科学院植物研究所专家确认标引的知识单元。人工标引主要从三个层次展开，如图 1 所示。

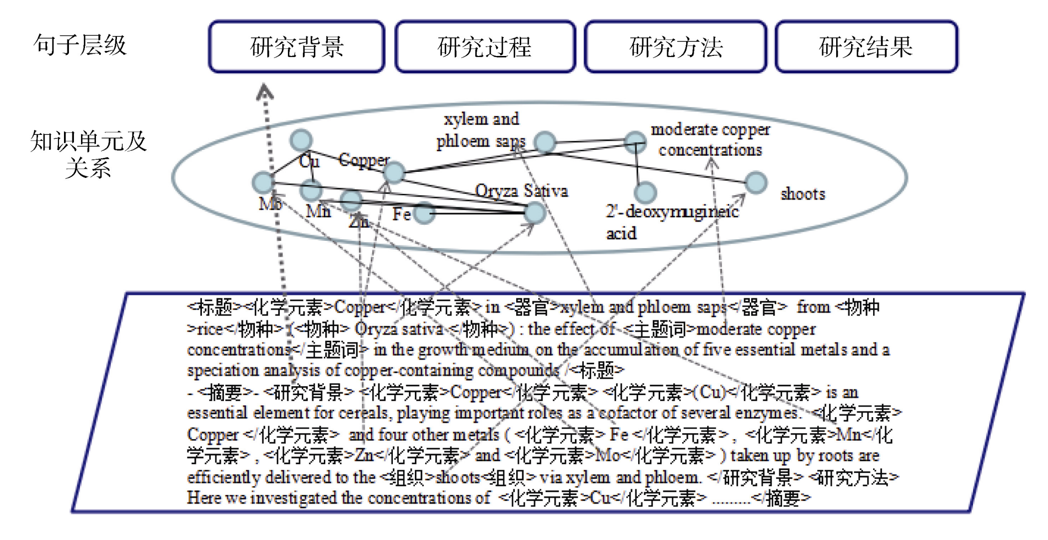


图 1　人工标引的层级示例

(1) 在句子层级，重点是通过标识特定用途的知识句群实现科技文献的结构化组织。在科技文献中，有不少知识无法简单地以某个知识单元(词组)或知识单元间关联关系的形式展示出来，比如一个完整的实验条件(化学元素的浓度与温度控制综合作用的实验条件)、一个完整的实验过程等，这些内容中可能包含多个知识单元和知识单元间的关联关系。针对此类信息，通过抽取识别关联密切的多个短语或短句形成知识句群，并确定各知识句群所属的类型，如研究方法、实验过程、研究结果等，可实现科技文献知识的重组织。

(2) 在知识单元层级，主要是考虑在实际的科技文献中，经常包含众多有明确的语义类别的知识单元，它们往往以命名实体名称或短语的形式在文献中呈现，承载了科技文献的主要知识内容。对这一类的知识单元进行抽取识别，可以从内容上实现科技文献内容的细粒度揭示，对后续的语义检索有重要意义。

(3) 在知识单元的关系层级，主要是考虑到知识单元并不是以独立、分散的形式存在于科技文献中，它们彼此之间往往还通过共现、主谓宾等表达方式，形成各种语义关联，结合这些语义关联可最大化地实现深层的文本内容挖掘。

上述三个层级的标引内容中，因为句子层次的抽取研究相对独立，笔者在前期论文中[17]已有专门论述，本文不再赘述。下文将重点详细论述知识单元及其之间的关系的具体内容。

3.2　语义知识框架的内容

在本研究的语义知识框架中，知识单元及知识单元之间的关系是重要内容。参照人工标引的结果和当前生物多样性领域重点关注的抽取内容，并结合项目的实际需求及后续抽取识别的可能性，笔者设计了如图 2 所示的植物物种多样性语义知识框架。该框架知识单元(图 2 中方框所示)以物种为核心，延伸了与物种相关的各类知识单元。其中，针对植物物种的属性描述，复用了植物本体(PO)①中部分概念，这些知识单元基本上涵盖了当前植物物种多样性科技文献中的主要知识点。这些知识单元之间除了上下位类的关联关系(图 2 中有箭头指向的知识单元的联系)外，不同类别的知识单元间也存在关联关系(图 2 中无箭头指向的知识单元间的联系)，通过共现、语法、语义等分析，可以构建形成这些知识单元之间的事实三元组，从而支持进一步的文本分析。

①美国国家科学基金会(NSF)资助构建的植物本体，是植物结构和生长阶段可控词汇表。

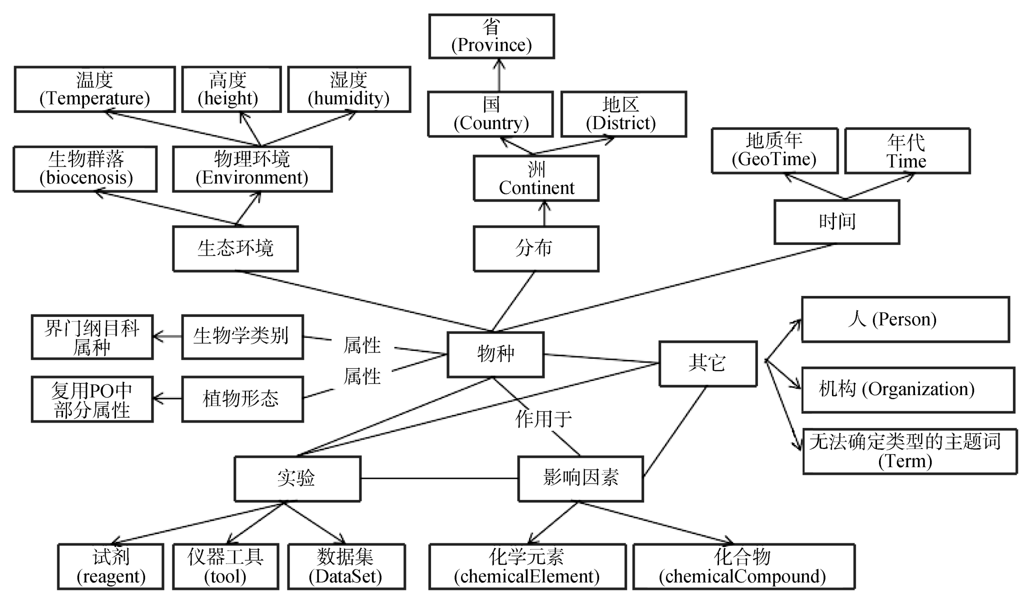


图 2 植物物种多样性语义知识框架

4 语义知识抽取的实现

植物物种多样性语义知识框架的构建一方面便于收集整理、组织已有的结构化知识，将现有的植物本体数据库等数据库中的记录作为图 2 中各类知识的实例进行存储，另一方面也为进一步的知识抽取提供了明确的目标。在此基础上，按照下述流程开展了植物物种多样性领域的知识抽取。

4.1 语料的整合与实验数据的选择

在确定植物物种多样性语义知识组织框架的基础上，通过专家咨询及参考中国科学院植物研究所的相关研究[18]，整理汇集了 G2000①植物本体数据库、NCBI 物种库[5]等相关领域术语和词汇、地址名称词表、Chemical Entities of Biological Interes②中的小化合物名称等语料，参照语义组织框架中定义的知识单元进行实例的整合，最终整合形成近 17 万条实例数据。这些领域资源一方面可作为词表直接用于知识单元实例的标注，另一方面，基于这些资源可通过半人工的方式构建实体识别规则库，用于新实例的识别。

此外，从 PubMed 中 Plant Physiology、The Plant Cell 两个期刊上获取了 23 000 篇期刊文摘，并根据中国科学院植物研究所提供的 20 种核心期刊列表，从 Web of Science 获取了 27 049 条科技文摘数据，构建了 5 万余篇实验数据集合。基于这些数据开展具体的知识抽取实验。

4.2 知识抽取框架的设计

为了更好地实现知识单元及知识单元间关系的识别，笔者设计了如图 3 所示的知识抽取框架。

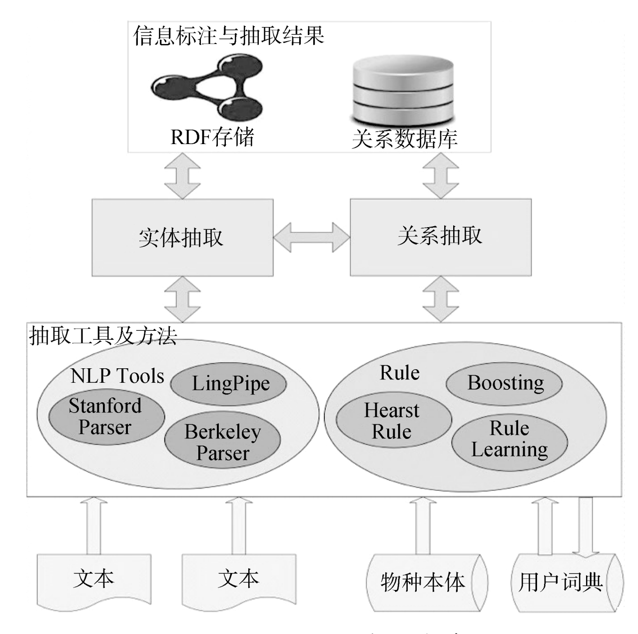


图 3 语义知识抽取框架

①一个植物物种本体库，由中国科学院植物研究所提供。

②Chemical Entities of Biological Interest(ChEBI)是可免费获取使用的一个生物化学实体本体，该本体集中关注小分子的化合物。

(1) 输入数据源：包括待抽取的科技文献及相关领域资源(植物多样性本体、NCBI 物种库等)。

(2) 抽取工具及方法：通过采用不同的自然语言处理工具(包括 Stanford Parser、Berkeley Parser 等)，实现对文本的词性标注、句法依存关系分析及句子的语法语义分析。

(3) 实体抽取与关系抽取：实体(即知识单元实例)抽取与关系抽取之间是交叉迭代实现的过程。一方面，实体抽取本身是一个迭代过程，新识别的命名实体添加到用户词典中，可用于下一轮的实体识别过程；另一方面，关系抽取的结果也可以用于发现新的实体，新发现的实体用于下一轮的关系发现过程。

(4) 信息抽取结果存储：根据信息抽取结果类型的不同，采用 RDF 存储和数据库存储两种方式实现实体及关系的存储。

4.3　知识单元实例及关系抽取

为了实现知识单元实例及关系的快速准确抽取，利用词典、规则、句法分析等方法开展抽取工作。其中，直接基于领域词典的实体标注是所有知识抽取研究的基础，本研究主要依赖词典实现部分物种名称、地理位置、部分化学元素与化合物、部分领域主题词等的抽取，具体过程与现有的相关研究并无差异，不再赘述，本文将重点论述基于规则的实例抽取与新实例的识别方法。

(1) 知识单元实例的标注与抽取

为了实现除词典中所含实例之外的知识单元实例的识别，笔者主要设计了以词典为基础，基于规则和统计方法相结合的方法。具体包含以下几个流程：

①基于规则的知识单元识别。尽管科技文献文本中的知识单元实例具体表现形式各不相同，但通过笔者对相关语料的分析，发现其组成词在词形、词性、组合方式等方面具有一些共性，如人物、机构、数值型信息、仪器设备等。针对这一类型的知识单元实例可通过人工辅助撰写规则的方式有效提高识别的准确率。针对此，笔者探索了如下快速构建规则的通用流程。

1)收集某一特定类别的知识单元实例样本，对该样本进行分词、分句、词性标注等自然语言处理。

2)针对组成比较简单的知识单元实例，如年份、日期、实验数据，以及相关的描述数值等，这类对象的识别可以借助构词法规则构建相关模式。

3)去除第 1)步分词结果中的介词、副词等无实际含义的词，从词频角度判断是否存在特殊的专有词汇(即类别指征词)。

4)针对有类别指征词的知识单元实例，在自然语言处理标注的基础上，将每个实例条目表现为词性、词形模式，其中特征词和名词以外词性的词保持原字符串输出。如图 4 两个示例，其中 Token 代表分词，Token.orth 表示该分词的拼字法，Token.category 表示词性。统计上述样本模式中特征词的位置，按照特征词位置对样本条目进行第一次分类，即分为特征词在“头部、中部、尾部”三类。在各个类别中，再按照词形组合进行分类，如不含有介词、所有格等词串的为一类，包含介词的词串为一类，包含所有格的为一类，均包含的为一类。根据第二次分类的结果，最终可以获取有效的模式组合。以高校为例，通过上述的分类学习，最终形成的高校实例模式基本包括：“<特征词> of NN/NNS*”(NN/NNS 标识首字母大写或全部大写的名词或名词复数，*代表多个 NN/NNS,)、“NN/NNS* <特征词>”、“(NN/NNS)('s)NN/NNS* <特征词>”、“NN/ NNS* <特征词> <介词> NN/NNS*”等。将以上学习出的模式转换为有限状态机，可用于实例识别的实现。

例1：University of New South Wales Australia
{(Token.string == University) (Token.string == of) (Token.orth == upperInitial, Token.category == NNP)*4 }

例2：Toronto's York University
{(Token.orth == upperInitial, Token.category == NNP) (Token.string =='s)
(Token.orth == upperInitial, Token.category == NNP) (Token.string == University)}

图 4　知识单元实例样本的模式输出示例

5)针对无类别指征词的知识单元实例，收集包含某类型科研要素实例的样本语句，人工标记出其中的科研要素实例作为训练样本；对所有的样本语句进行分词、词性标注、句法解析等操作，获取相关的语言学特征；获取样本语句中科研要素实例的 n 个上下文临近词(n 可以灵活调整，本研究中参考 Jiang 等[19]的研究，将 n 设定为 4)，统计这些上下文临近词的词频，获取前三个最高词频的临近词，统计包含这些高频临近词的词条百分比，若超出 50%，可以认为该词具有此类科研要素前导词或后引词的语义特征。对上文获取的前导词或后引词分别从 WordNet 中获取其同义词集合 *Synset*[*train*]，逐个解析待识别科研要素所在的句子，同样获取其 n 个上下文临近词，分别获取这些临近词在 WordNet 中的同义词集合 *Synset*[*test*]，计算 n 个临近词的 *Synset*[*train*] 与 *Synset*[*test*]相似度之和，见公式(1)和公式(2)。这里选择使用计算同义词集合的最大相似度替代直接的词与词之间的相似度，主要考虑到现实文本中，各语义在用词选择、词的词形、拼写等方面存在变异等情况，计算同义词集合的最大相似度可以降低这些情况造成的词间相似度过低的影响。

$$Sim = \sum Sim(nw) \tag{1}$$

$$Sim(nw) = Sim(Synset_{train}, Synset_{test}) \tag{2}$$

其中，*Sim* 是最终的总体相似度，*Sim*(*nw*)是每一个临近

词的相似度。

通过上述的规则，一方面可以直接实现知识单元实例的抽取识别，另一方面，还可获取候选实例，以便进一步通过其他方法确认其语义类别。但是依赖于人工辅助撰写规则库虽然准确性较高，规则覆盖的全面性会直接影响抽取的查全率。因此，还有一部分知识单元实例需要通过其他方法来识别。

②基于词典相似性的知识单元实例识别。虽然基于词典的实例识别无法解决新实例的识别问题，但词典依然为新实例的识别提供了重要的支撑。依据词形、词性、词频、句法成分等特征，可遴选出待抽取的新知识单元实例集合。然后进一步计算候选实例与词典中实例之间的编辑距离，获得候选知识单元与词典中实例的相似度，以实现对一些未登录词的识别。

③基于句法分析的知识单元实例识别。知识单元实例抽取与关系抽取之间是一个交叉迭代实现的过程，关系抽取的结果可用于发现新的实例。针对一些通过规则、词典相似性仍无法识别的候选实例，可以借助于句法分析中获得的句法依存关系及语法关系(并列的句子成分)，结合统计分析的算法，实现实例语义类型的判别。

具体而言，借助 Parser 进行句法解析时，可将句子表示为层级的句法树。以“Bell, based in Los Angeles, makes and distributes electronic, computer and building products.”为例，其经过 Parser 解析的句法树如图 5 所示，其句法标记采用了 Penn 树库[20]，这与多数词性标注系统都可兼容。

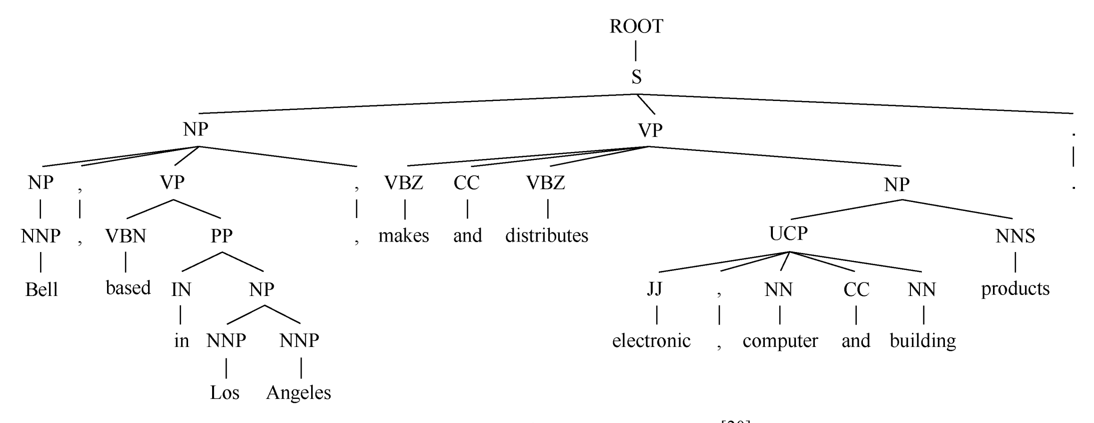


图 5　Parser 解析生成的句法树[20]

除句法解析树外，还可以借助 Parser 获取句法依存分析结果，如图 6 所示。图中右半边的依存句法分析结果说明中，括号内的都是句中的实例，括号前的 nsubj、partmod 等关键词标识特定的依存关系。

例句：Bell, based in Los Angeles, makes and distributes electronic, computer and building products.

依存句法分析结果：
nsubj（makes-8，Bell-1）；
nsubj（distributes-10，Bell-1）；
Partmod（Bell-1，based-3）；
nn(Angeles-6,Los-5);
prep_in(based-3, Angeles-6);
conj_and(makes-8, distributes-10);
amod(products-16，electronic-11);
conj_and(electronic-11，computer-13);
amod(products-16，computer-13);
conj_and(electronic-11，building-15);
amod(products-16，building-15);
dobj（makes-8, products-16）；
dobj(distributes-10, products-16)

依存句法分析结果说明：
nsubj（makes-8，Bell-1）；nsubj（distributes-10，Bell-1）；
表明Bell与makes、distributes的主谓关系；
Partmod（Bell-1，based-3）；表明以based动词分词形式修饰Bell；
nn(Angeles-6,Los-5); 表明Los Angeles为复合名词词组；
prep_in(based-3, Angeles-6); 表明介词关系；
conj_and(makes-8, distributes-10); 表明makes与distributes并列关系；
amod(products-16，electronic-11); amod(products-16，computer-13);
amod(products-16，building-15);
表明products中包含了electronic、computer和building，
conj_and(electronic-11，computer-13);conj_and(electronic-11，building-15);
表明electronic、computer和building的并列关系；
dobj（makes-8, products-16）;dobj(distributes-10, products-16)
表明products是makes、distributes的直接宾语

图 6　Parser 解析生成的依存语法结果及说明

从图 6 中可以看到，依据句法分析，可以比较清晰地获取复合名词短语，即句法树中的NP模块，而复合名词短语内各词的依存关系特征也可以清楚地显示出词组内的关联关系。如通过 conj_and(electronic-11, computer-13)可以获取到 electronic 和 computer 的并列关系，若确定了这两者中任意一个的语义类别，即可判定另外一个的类别，从而实现实例类型的标注。

(2) 关系的抽取

知识单元之间的关系存在很多种，包括共现、同位语法、并列语法、事实关系、语义上下位类等，其中共现关系最为简单，依据两个知识单元实例是否共同出现在指定的窗口内(全文、摘要、句子)，即可判定二者之间是否存在共现关系，这也是判定知识单元之间是否存在相关性的最简单直接的做法。本研究中因为处理的文本对象为期刊的摘要，因此，在构建共现关系时，选用句子作为共现窗口，即如果两个实例共同出现在同一个句子中，即为共现关系。此类关系的判别比较简单，本文不做详细说明，而是重点介绍基于句法分析的语法、事实与语义规则的关系识别。

①同位与并列语法关系抽取。如上文"基于句法分析的知识单元实例识别"中所描述，有一类新实例的识别即借助于实体的同位与并列语法的关系来鉴定。同样，在确认了两个实例的类型之后，基于句法解析和句法依存关系解析中所获取的 and、or 等关系，即可确认两个实例之间的同位与并列语法关联。

②事实关系识别。本文所论及的事实关系主要是指在文中存在的主谓宾关系，即<S, P, O>(主语，谓词，宾语)事实，这一类关系可为后续的推理提供重要的支持。针对此类关系的识别，笔者设计如下流程:

1)输入已完成分词、分句及知识单元实例识别的文本。以分句结果为循环处理参照，逐个处理每个分句。针对非动词谓词和动词谓词分别构建两个空的关系三元组列表。

2)判断每个分句中是否包含一个以上图 2 中定义的知识单元实例。若不包含，则结束此句分析，返回步骤 1)，继续下一个句子的分析；若包含两个及两个以上的科研要素，继续步骤 3)。

3)依照 Parser 解析器解析的句法结果，自句法树底层开始，逐层构建最简结构的简单句(即仅包含一个主谓宾结构的句子，而不包含任何的从句)，构建分句的简单句群。此步骤获得的简单句群成为第二个循环点。

4)逐个判断简单句群中每一个简单句是否包含一个以上图 2 中定义的知识单元实例。若不包含，则结束此句分析，返回继续下一个句子的分析；若包含，继续步骤 5)。

5)借助 Parser 解析器获取的句法依存关系，获取主谓宾关系，构建(主语词组，谓语动词，宾语词组)的关系三元组。继续步骤 6)。

6)进一步分析(主语词组，谓语动词，宾语词组)的关系三元组，判断在主语词组和宾语词组部分是否均存在至少一个图 2 中定义的知识单元实例。若有，进入步骤 7)；若所有实例全部存在于同一个词组部分，则跳到步骤 8)。

7)若主语词组和宾语词组中均仅存在一个实例，判断这些科研要素实例间是否存在转义问题。若无，则构建相应的关系三元组，添加入动词关系三元组列表中。若存在转义，则依据转义语义关系选择是否放弃该关系三元组。若主语词组和宾语词组中存在一个以上的实例，则基于排列组合的方法依次处理，但是需要注意，此时需要注意并列问题所引起的歧义。

8)分析相应词组中的科研要素实例，借助其标注类型等信息，判定其语义关系。

9)输出动词关系三元组列表。

③语义上下位关系的发现。这一类关系主要是以所有格、固定句式、常用表达(如 such as, for example, as well as 等)为代表的有限关系，这类关系可以主要通过模式规则隐式构建关系三元组，发现实例之间的语义上下位关系。为了实现此类关系的识别，笔者主要参考 Hearst 模式[21]进行相关模式规则的扩展，人工构造了 20 多条关系规则，从而实现了相应关系的识别。

4.4　知识抽取的结果应用

基于上述的知识抽取方法，从 5 万多篇相关的文献标题和摘要中共获得 273 668 条知识单元的实例抽取结果，主要的抽取类型分布结果如表 1 所示(只展示了抽取实例数量大于 100 的结果)。

表 1　从实验数据中抽取的主要知识单元及物种属性实例分布

实体类型	数量	实体类型	数量
物种-属(Genus)	115 698	植物茎(plantStemForm)	1 983
物种-科(family)	25 332	省(province)	1 845
习性(habit)	13 510	花期(plantFlowerTime)	1 773
花颜色(plantFlowerColor)	12 649	植物根类型(plantRootType)	1 725

(续表)

实体类型	数量	实体类型	数量
生态环境(cultivatedHabitat)	12 277	化合物(chemicalCompound)	1 637
植物茎类型(plantStemType)	10 306	授粉系统(plantPollinationSystem)	1 509
物种-种(species)	9 478	基因(gene)	1 270
寿命(longevity)	8 233	国家(country)	1 227
植物果实类型(plantFruitType)	6 489	物种-目(order)	1 088
植物雌蕊融合(plantGynoeciumCarpelFusion)	4 875	花对称性(plantFlowerSymmetry)	1 043
植物雄蕊排列(planAndroeciumStamenArrangement)	4 793	化学元素(ChemicalElement)	736
植物叶规格(plantLeafArrangement)	3 908	实验材料与工具(Tool)	722
植物叶形状(plantLeafShape)	3 609	物理环境(PhysicalEnvironment)	717
植物叶缘(plantLeafMargin)	3 268	植物花被(plantFlowerPerianthForm)	621
花序形态(plantInflorescenceForm)	3 268	植物叶结构(plantLeafStructure)	510
植物叶面(plantLeafSurface)	2 859	器官(Organ)	780
花结构数量(plantNumbersOfFloralStructure)	2 815	机构(Organization)	323
植物叶部(plantLeafDivision)	2 615	培养环境(culturedHabitat)	264
无确定类型的主题词(Term)	2 482	物种-门(phylum)	252
光合作用(photosynthesis)	2 282	植物叶(plantLeaf)	244
植物叶性(plantFlowerSexuality)	2 222	物种-纲(class)	153
植物雄蕊类型(plantAndroeciumStamenType)	2 152	植物根结构(plantRootStructure)	127

	identifier	s	p	o	literalIdentifier	content	score	zone	annotatedTime
1	b4de28adcb3522939...	Allantodia , Callipteris and Monomelangium	nested within	Diplazium	1177899	Phylogenetic analyse...	0.886428672103566	Abstract_En	2014-05-13 17:03
2	df502c14cabb721c5...	eight robust sub-clades	were found in	the phylogenetic topology	1177899	Four well-supported ...	0.878434364515284	Abstract_En	2014-05-13 17:03
3	9963e922b3073abde...	Bayesian methods	congruently resolved	Allantodia , Callipteris and Monomelangium	1177899	Phylogenetic analyse...	0.488757640468295	Abstract_En	2014-05-13 17:03
4	9a4d2e0a634bff25de...	we	sampled over	6000 DNA nucleotides of up to seven plas...	1177899	For each species, w...	0.628467956197211	Abstract_En	2014-05-13 17:03
5	07000c8e26782e842...	the phylogenetic relationships of these ge...	were investigated using	a comprehensive taxonomic sampling	1177899	In the present study, ...	0.418249052182681	Abstract_En	2014-05-13 17:03
6	8b370ebd71988f9bc...	petiole/rachis scales	recovered	some character combinations of systemati...	1177899	Reconstruction of th...	0.707866114109794	Abstract_En	2014-05-13 17:03
7	3b7fcc210cdf50196f...	Diplazium and allied segregates (Allantod...	represent highly	diverse genera	1177899	Diplazium and allied ...	0.365868557906458	Abstract_En	2014-05-13 17:03
8	ce52729b91c1104bb...	The genetic differentiation pattern	was reflected by	morphological differentiation	1177900	The genetic differenti...	0.975776075417946	Abstract_En	2014-05-13 17:03
9	43fb41634639c81a9...	octoploids	suggests an autopolyploid origin of	the latter	1177900	Lack of AFLP diverg...	0.58619014167641	Abstract_En	2014-05-13 17:03
10	60b10afe4adfde9ed4...	some of the previously described taxa	constitute	distinct genetic entities	1177900	While some of the pr...	0.707738196724996	Abstract_En	2014-05-13 17:03
11	63811d8181cd063af...	Amplified Fragment Length Polymorphism...	revealed	four clearly differentiated genetic groups	1177900	Eur. based on Amplifi...	0.0322806084744039	Abstract_En	2014-05-13 17:03
12	7c4b22240f9a85bbe...	others	have	no taxonomic value	1177900	While some of the pr...	0.572506043709676	Abstract_En	2014-05-13 17:03
13	c2e19115011d69b80...	four clearly differentiated genetic groups	did only partly follow	recent taxonomic concepts	1177900	Eur. based on Amplifi...	0.168288304874133	Abstract_En	2014-05-13 17:03
14	8d5ef329878872a6d...	octoploids	occur only within	S. filifolia	1177900	Synthesizing our AFL...	0.681725466799967	Abstract_En	2014-05-13 17:03
15	f36f2f716ef2f292534...	Cannabaceae	includes	ten genera	1177901	Cannabaceae includ...	0.216087023726981	Abstract_En	2014-05-13 17:03
16	7ecce58a6c1f18483...	ten genera	are widely distributed in	tropical	1177901	Cannabaceae includ...	0.123284121924827	Abstract_En	2014-05-13 17:03
17	4a2e4530e42a71f5e...	Trema	was paraphyletic with	respect	1177901	All genera were mon...	0.255245059288544	Abstract_En	2014-05-13 17:03
18	2b2816d649e8da281...	All genera	were monophyletic except for	Trema	1177901	All genera were mon...	0.986271693259343	Abstract_En	2014-05-13 17:03
19	0730d4931d073140e...	The molecular results	strongly supported	this expanded family	1177901	The molecular results...	0.426901174096539	Abstract_En	2014-05-13 17:03
20	69c8f00c0c2c3d40e...	a strongly supported clade	further resolved into	a Lozanella clade	1177901	The Aphananthe cla...	0.963702237941468	Abstract_En	2014-05-13 17:03

图 7　植物物种多样性领域 SPO 语法关系部分抽取结果

除表 1 所示的知识单元的实例外，本研究从实验数据中抽取获得 133 922 条 SPO 语法关系，抽取获得 35 903 条同位语关系结果。图 7 展示了 SPO 语法关系的部分抽取结果。

基于上述知识抽取的结果，综合利用领域知识库和其他第三方资源，进一步构建了植物物种多样性领域语义检索的应用示范平台，为用户提供领域知识揭示、语义标注、本体导航等检索应用，验证了本研究成果的有用性和有效性。图 8、图 9 分别展示了部分的应用示范。

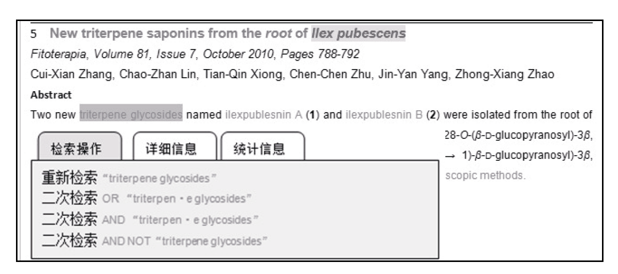


图 8　基于本体概念或实体的知识浏览、检索与统计分析功能

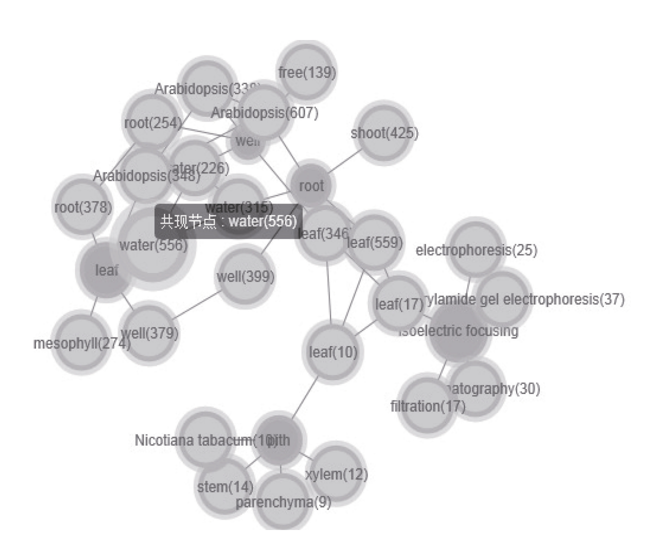


图 9 基于语义知识抽取的单篇文章共现关系知识图

5 结 语

与一般的生物知识抽取相比，植物物种多样性领域涉及的知识单元类型及其关系更为复杂，如生态环境、物种特征、影响因素等，因此，在设计植物物种多样性领域语义知识框架时需要从最终的应用角度考虑更多知识单元，在具体识别中，需要针对不同类别的知识单元综合采用更多领域无关的知识抽取方法，以便适应多类知识单元实例的抽取识别。

在对当前生物多样性信息抽取领域相关研究分析的基础上，本文结合中国科学院文献情报中心"建设生物多样性领域本体构建与语义组织应用示范平台"的实际要求，设计了植物物种多样性语义知识抽取框架，探索实现了相应的语义知识抽取方法。本研究更多从实际应用的层面探索了可工程化应用的知识组织框架及知识识别的方法，因此，词典和人工撰写的规则是本研究中开展知识抽取的重要组成部分，正因为此，词典和人工规则本身所固有的局限性也在一定程度上限制了识别的完整性和准确性，在未来，针对各类型知识单元的精细化识别仍将是重要研究内容。

参考文献：

[1] Thessen A E, Cui H, Mozzherin D. Applications of Natural Language Processing in Biodiversity Science [J]. Advances in Bioinformatics, 2012. DOI: 10.1155/2012/391574.

[2] Naderi N, Kappler T, Baker C J, et al. OrganismTagger: Detection, Normalization and Grounding of Organism Entities in Biomedical Documents [J]. Bioinformatics, 2011, 27(19): 2721-2729.

[3] Species [EB/OL]. [2016-04-12]. http://en.wikipedia.org/wiki/Species.

[4] Gerner M, Nenadic G, Bergman C M. LINNAEUS: A Species Name Identification System for Biomedical Literature [J]. BMC Bioinformatics, 2010. DOI: 10.1186/1471-2105-11-85.

[5] The NCBI Taxonomy Homepage [EB/OL]. [2016-04-12]. http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/.

[6] Page R D M. BioNames: Linking Taxonomy, Texts, and Trees [OL]. http://dx.doi.org/10.7717/peerj.190.

[7] Species 2000 [EB/OL]. [2016-04-12]. http://www.catalogueoflife.org/annual-checklist/2014/.

[8] Akella L M, Norton C N, Miller H. NetiNeti: D1iscovery of Scientific Names from Text Using Machine Learning Methods [J]. BMC Bioinformatics, 2012. DOI: 10.1186/1471-2105-13-211.

[9] The OrganismTagger System [EB/OL]. [2016-04-12]. http://www.semanticsoftware.info/organism-tagger.

[10] Koning D, Sarlar I N, Moritz T.Taxongrab: Extracting Taxonomic Names from Text [J]. Biodiversity Informatics, 2005, 2: 79-82.

[11] Taylor A. Extracting Knowledge from Biological Descriptions [C]//Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases. 1995: 114-119.

[12] Tang X, Heidorn P B. Using Automatically Extracted Information in Species Page Retrieval [C]//Proceedings of TDWG 2007. 2007.

[13] Cui H. CharaParser for Fine-grained Semantic Annotation of Organism Morphological Descriptions [J]. Journal of the Society for Information Science and Technology, 2012, 63(4): 738-754.

[14] 段宇锋，黄思思. 中文植物物种多样性描述文本的信息抽取研究[J]. 现代图书情报技术，2016(1): 87-96. (Duan Yufeng, Huang Sisi. Information Extraction from Chinese Plant Species Diversity Description Text[J]. New Technology of Library and Information Service, 2016(1): 87-96.)

[15] Li C, Liakata M, Rebholz-Schuhmann D. Biological Network Extraction from Scientific Literature: State of the Art and Challenges [J]. Briefings in Bioinformatics, 2013. DOI: 10.1093/bib/bbt006.

[16] Skusa A, Rüegg A, Köhler J. Extraction of Biological Interaction Networks from Scientific Literature[J]. Briefings in Bioinformatics, 2005, 6(3): 263-276.

[17] 白光祖，何远标，马建霞，等. 利用小样本量机器学习实

现学术文摘结构的自动识别[J]. 现代图书情报技术, 2014(7-8): 34-40. (Bai Guangzu, He Yuanbiao, Ma Jianxia, et al. Application of Machine Learning with Limited Corpus to Identify Structure of Scientific Abstracts Automatically, 2014 (7-8): 34-40.)

[18] 许哲平, 崔金钟, 覃海宁, 等. 中国植物物种多样性 e-Science 平台建设构想[J]. 植物物种多样性, 2010, 18(5): 480-488. (Xu Zheping, Cui Jinzhong, Qin Haining, et al. On the Architecture of Biodiversity e-Science Infrastructure in China[J]. Biodiversity Science, 2010, 18(5): 480-488.)

[19] Jiang W, Guan Y, Wang X L. Improving Feature Extraction in Named Entity Recognition Based on Maximum Entropy Model [C]//Proceedings of the 5th International Conference on Machine Learning and Cybernetics. 2006: 2630-2635.

[20] De Marneffe M-C, Manning C D. Stanford Typed Dependencies Manual [OL]. http://nlp.stanford.edu/software/dependencies_manual.pdf.

[21] Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora [C]// Proceedings of the 14th International Conference on Computational Linguistics, 1992.

作者贡献声明:

刘建华: 提出论文整体框架, 完成语义知识抽取框架的设计, 参与实现知识抽取的实现开发, 完成论文主体内容的撰写, 对最终版本部分内容进行校对、修改完善;

王颖: 参与设计语义知识抽取框架和知识抽取开发的语料准备、存储结构设计、数据处理;

张智雄: 参与语义知识抽取框架的设计, 对论文内容提出修改意见;

李传席: 负责知识抽取功能的实现开发, 并提供开发文档。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: liujh@mail.las.ac.cn。

[1] 刘建华, 王颖, 李传席. Top40000 条 SPO 抽取结果.xls.

收稿日期: 2016-04-14
收修改稿日期: 2016-08-12

Extracting Semantic Knowledge from Plant Species Diversity Collections

Liu Jianhua[1, 2] Wang Ying[1] Zhang Zhixiong[1] Li Chuanxi[3]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](China Great Wall Asset Management Co., Ltd, Beijing 100045, China)

Abstract: [**Objective**] This paper aims to extract semantic knowledge from the biodiversity studies. [**Methods**] We proposed a new knowledge extraction framework focusing on species. It included various entities as well as the relationship among them. The new method was then examined with various specialized databases. [**Results**] The species-oriented knowledge extraction framework, could successfully retrieve semantic information from the target entities and the relations among them. This method expanded the scope of knowledge extraction practice in the biodiversity field. [**Limitations**] The recall and precision ratio of the new method was effected by the dictionaries and rules. More studies are needed to examine the semantic relationship among the named entities beyond co-occurrence, hierarchical and simple syntactic relations. [**Conclusions**] The proposed method expands the contents and methods of knowledge extraction in biodiversity research. It supports the semantic information retrieval and computation.

Keywords: Plant Species Diversity Plant Species Knowledge Extraction Relation Extraction